Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8403–8407, August 1995
Genetics

Sequence analysis of the cis-regulatory regions of the bithorax complex of *Drosophila*

(Ultrabithorax/abdominal-A/Abdominal-B/Markov chain/cis regulation)

EDWARD B. LEWIS, JOHN D. KNAFELS, DAVID R. MATHOG, AND SUSAN E. CELNIKER

Division of Biology, California Institute of Technology, Pasadena, CA 91125

Contributed by Edward B. Lewis, June 19, 1995

ABSTRACT **The bithorax complex (BX-C) of *Drosophila*, one of two complexes that act as master regulators of the body plan of the fly, has now been entirely sequenced and comprises ≈315,000 bp, only 1.4% of which codes for protein. Analysis of this sequence reveals significantly overrepresented DNA motifs of unknown, as well as known, functions in the non-protein-coding portion of the sequence. The following types of motifs in that portion are analyzed: (*i*) concatamers of mono-, di-, and trinucleotides; (*ii*) tightly clustered hexanucleotides (spaced ≤5 bases apart); (*iii*) direct and reverse repeats longer than 20 bp; and (*iv*) a number of motifs known from biochemical studies to play a role in the regulation of the BX-C. The hexanucleotide AGATAC is remarkably overrepresented and is surmised to play a role in chromosome pairing. The positions of sites of highly overrepresented motifs are plotted for those that occur at more than five sites in the sequence, when <0.5 case is expected. Expected values are based on a third-order Markov chain, which is the optimal order for representing the BXCALL sequence.**

The bithorax complex (BX-C) is a set of master control genes that programs the development of the abdomen and a portion of the thorax of the fruit fly (reviewed in refs. 1 and 2). The complex consists of three homeobox-containing genes—Ultrabithorax (*Ubx*), abdominal-A (*abd-A*), and Abdominal-B (*Abd-B*)—and 12 cis-regulatory regions. The anterobithorax (*abx*), bithorax (*bx*), and postbithorax (*pbx*) regions function in the wild type to promote development of the third thoracic segment into a haltere-bearing, instead of a wing-bearing, segment. The remaining nine regions, *bxd* and *iab-2–iab-9*, inclusive, determine the pattern of differentiation of abdominal segments A1–A9, respectively.

The complex has been sequenced and comprises ~315,000 bases. The methods used in sequencing it and an analysis of its open reading frames are described in the preceding paper (3). The sequence itself is highly unusual in that >98% of it fails to code for protein. The challenge is to identify in the noncoding fraction DNA sequence motifs that are involved in cis regulation of the complex, such as activators or repressors of transcriptional initiation, enhancers or silencers of transcription, and motifs involved in such processes as DNA replication, splicing, chromatin activation and deactivation, and chromosome pairing. Also expected are binding sites for trans-acting repressor proteins of the Polycomb (*Pc*) family type (reviewed in ref. 4) and trans-acting activator proteins of the trithorax (*trx*) type (reviewed in ref. 5). Finally, an unknown fraction of the noncoding sequence may consist of spacer DNA that is needed to establish correct boundaries for proper expression of motifs, especially those that function in a clustered fashion.

This paper reports a preliminary analysis of DNA motifs that for the most part are highly overrepresented, singly and/or in a clustered manner. In some cases, biochemical and developmental studies have already identified motifs that have functional significance; hence their degree of abundance becomes of interest. In other cases, there are sequences so overrepresented as to suggest that they also have functional importance. The latter cases, if verified by molecular and developmental studies, will extend the known repertoire of motifs thought to regulate the complex. Increasingly, as sequence analysis becomes available for noncoding regions of the complex in other organisms, such as *Drosophila virilis*, the degree of sequence conservation of motifs between species will help identify those that are functionally indispensable.

MATERIALS AND METHODS

The BX-C is located in the 89E region of the salivary gland chromosomes. A total of 338,324 bp from that region have now been entirely sequenced and designated SEQ89E (3) (GenBank accession no. U31961). At each end of SEQ89E are sequences of putative genes that appear to be functionally unrelated to the BX-C. We have therefore deleted such genes to generate a sequence of 314,895 bp, designated BXCALL, that is expected to include the entire BX-C. All base-pair positions in this paper refer to SEQ89E.

Since the sequence of base pairs in noncoding as well as in coding regions of eukaryotic DNA is known to be nonrandom, Markov chain theory has been adopted to represent such sequences (reviewed in ref. 6). We find that a third-order Markov chain (TMC) best describes the sequence, based on the Bayesian independence criterion (BIC) (7). Thus, for all order tested from the zeroth—identical with base pair frequencies generated independently—to the fifth, the BIC is at a minimum for the third order (TMC) (data not shown).

The TMC assumes that the probability of occurrence of a given base in a sequence is not independent of the base that precedes it but instead is conditional upon the probability of obtaining the three bases that immediately precede it. As an example, the probability of obtaining AGATAC in BXCALL is the probability of obtaining the trinucleotide AGA multiplied by the three conditional probabilities of obtaining the three bases which follow it. An estimate of the conditional probability of obtaining a T following the trinucleotide AGA can be obtained as the ratio of the total number (n) of AGAT tetranucleotides observed in the entire sequence (BXCALL) to the total number of AGA trinucleotides in that sequence; similarly, for the A which follows GAT, the estimate becomes the ratio of the number of GATA tetranucleotides to the number of GAT trinucleotides, and for the C that follows ATA the estimate is the ratio of the number of ATAC tetranucleotides to the number of ATA trinucleotides. The expression for the probability (P) of obtaining the sequence AGATAC is

$$P(\mathrm{AGATAC}) = P(\mathrm{AGA})[P(\mathrm{T/AGA}) \times P(\mathrm{A/GAT}) \times P(\mathrm{C/ATA})],$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: TMC, third-order Markov chain.

which can be estimated as

$$(n_{AGAT} \times n_{GATA} \times n_{ATAC})/(n_{GAT} \times n_{ATA}).$$

To derive expected numbers of occurrences of a given clustered sequence, when a space is allowed between members of the cluster, a pseudorandom method is adopted. We used the observed frequency of a given clustered sequence in 100 pseudorandom control sequences based on the TMC to estimate its expected frequency in BXCALL. The program MAKE_RANDOM_DNA (devised by D.R.M.) was used to generate such control sequences. It sequentially assigns bases by using a pseudorandom number generator and the Markov transition weightings derived from BXCALL.

RESULTS

DNA Motifs of Unknown Function. We have identified (Fig. 1) significantly overrepresented concatamers of mono-, di-, and trinucleotides for which the TMC expectation in every case is <0.5 (Table 1). Overrepresented clusters of two or more hexanucleotides (spaced ≤5 nt apart) have been identified with the aid of the CORES program (devised by D.R.M.). Excluding rotational derivatives, we have identified six such clusters (Fig. 1) on the basis that for each cluster at least six cases occur in BXCALL, and the TMC expectation, in every case, is <0.5 case (Table 2). If the spacing between members of the cluster is allowed to increase and/or the number of

Table 1. Comparison of observed (O) and expected (E) numbers of mono-, di-, and trinucleotide concatemers in BXCALL

Motif	O	E-M*	E-I†	P‡
$(A)_{15}$	5	0.182	0.004	$<10^{-5}$
$(T)_{15}$	3	0.178	0.004	$<10^{-3}$
$(CT)_6$	4	0.022	0.018	$<10^{-8}$
$(AG)_6$	3	0.007	0.016	$<10^{-7}$
$(CA)_6$	14	0.063	0.018	$<10^{-15}$
$(TG)_6$	6	0.057	0.016	$<10^{-10}$
$(AT)_6$	4	0.214	0.140	$<10^{-4}$
$(CAG)_4$	5	0.104	0.008	$<10^{-7}$
$(CTG)_4$	4	0.064	0.008	$<10^{-6}$
$(ATA)_4$	2	0.204	0.127	0.018
$(TAT)_4$	4	0.213	0.125	$<10^{-4}$

*Expectation based on the TMC method.
†Expectation based on the independence of the single base frequencies.
‡Probability that the observed (O) number or a larger number exceeds E-M based on the cumulative Poisson distribution.

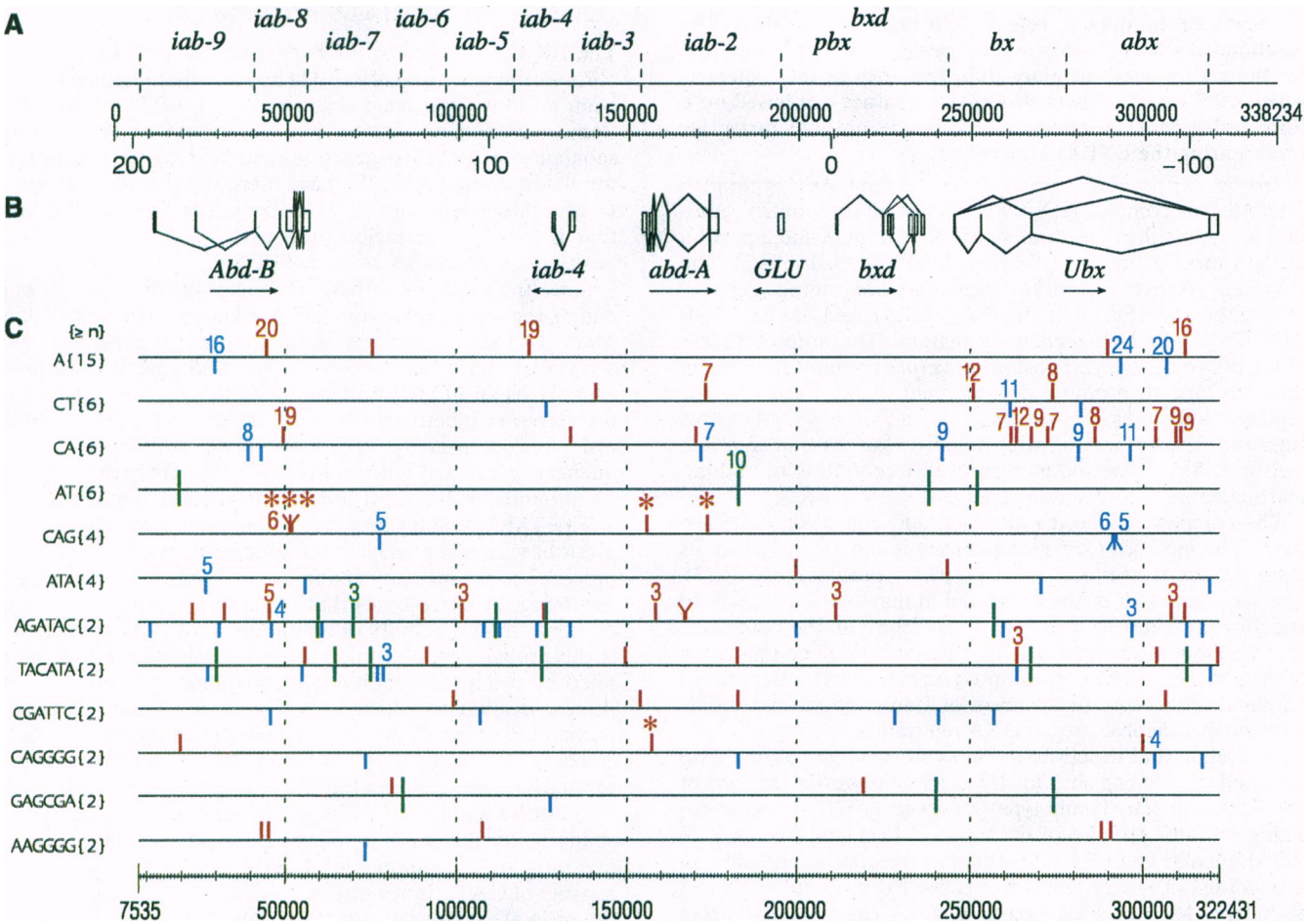

FIG. 1. Correlation of molecular and genetic maps of the BX-C. (*A*) Boundaries of cis-regulatory regions delineated by hatched vertical lines are only approximately known. Numbers below the line refer to the map positions in kilobases. The bases numbered 1 and 313,000 correspond approximately to +204 kb and −100 kb, respectively, on the BX-C walk (8, 9). (*B*) The transcription map is based on intron–exon mapping, cDNA, and computer analysis. Transcription units include homeobox-containing *Abd-B, abd-A*, and *Ubx*; the noncoding *bxd* and *iab-4*; and the predicted glucose transporter-like gene (3). Arrows indicate the direction of transcription. (*C*) Distribution of DNA sequence motifs. Motifs observed four or more times in BXCALL yet having a TMC expectation of 0.5 or less. The number above each plotted position refers to the number of repeats of the concatamer at that location. The position of concatamers of a given DNA motif (m) are shown in red; that of its complement (m′), in blue; and that of concatamers of both m and m′, in green.

Table 2. Comparison of observed (O) and expected (E) numbers of motifs of unknown function in BXCALL

Motif*	O	E-M†	E-I‡	P§
AGATAC	138	44.790	99.433	$<10^{-15}$
GTATCT	121	39.511	98.913	$<10^{-15}$
$(AGATAC)_{n\geq 2}$	31	0.12	0.752	$<10^{-15}$
TACATA	141	78.894	141.348	$<10^{-9}$
TATGTA	126	74.622	137.954	$<10^{-7}$
$(TACATA)_{n\geq 2}$	19	0.41	1.494	$<10^{-15}$
CGATTC	76	52.151	71.107	0.001
GAATCG	101	58.007	69.764	$<10^{-6}$
$(CGATTC)_2$	9	0.23	0.378	$<10^{-11}$
CAGGGG	57	48.357	36.433	NS
CCCCTG	61	52.667	36.652	NS
$(CAGGGG)_{n\geq 2}$	7	0.21	0.096	$<10^{-8}$
GAGCGA	80	56.262	49.077	0.002
TCGCTC	77	65.773	50.984	NS
$(GAGCGA)_2$	6	0.29	0.191	$<10^{-6}$
AAGGGG	83	66.317	48.024	0.027
CCCCTT	93	72.793	52.102	0.013
$(AAGGGG)_2$	6	0.40	0.191	$<10^{-5}$

*Each group of three lines shows a motif (m), its complement (m'), $(m)_2$ = m#m, m#m', m'#m, and m'#m', where # = spacing of ≤5 nt.

†Expectation based on the TMC method.

‡Expectation based on independence of the single base frequencies.

§Probability that the observed (O) number or a larger number exceeds E-M based on the cumulative Poisson distribution. NS, not significant ($P > 0.05$).

observed cases in BXCALL is <6, many more types of hexanucleotides are significantly overrepresented; however, since the TMC expectation exceeds 1, they are not plotted in Fig. 1.

A limited study has been made of repeats in BXCALL. Direct repeats were identified in two ways, using the REPEAT program of the Genetics Computer Group (41) or the BLAST program (42), which compares the sequence with itself. Reverse repeats were analyzed with the STEMLOOP program of the Genetics Computer Group or the BLAST program. Since the maximum length of a perfect repeat found in performing these programs on 10 of the TMC-generated sequences of BXCALL was 18 bp, we have searched only for perfect repeats of 20 or more. Only three such direct repeats were found in BXCALL after we excluded the long repeats of mono-, di-, tri-, and hexanucleotides already considered in Fig. 1. These three repeats map within the *Abd-B* domain: one of 110 bp at positions 68,556 and 68,656; one of 31 bp at 26,013 and 26,045; and one of 24 bp at 31,458 and 31,482. Only two reverse repeats of 20 bp or more were identified, one of 24 bp at positions 246,197 and 246,138 and one of 21 bp at positions 50,785 and 57,889.

Motifs Derived from Consensus DNA/Protein Binding Sites. Activation and/or repression of the BX-C genes (reviewed in ref. 10) has been inferred from analyses of mutants of the gap genes giant (*gt*), hunchback (*hb*), tailless (*tll*), knirps (*kni*) and Kruppel (*Kr*); the pair-rule gene fushi-tarazu (*ftz*); homeotic genes such as caudal (*cad*), *Ubx*, *abd-A*, and *Abd-b*; and the pairing-related gene zeste (*z*).

We have deduced consensus DNA-binding motifs from *in vitro* footprinting studies for the protein products of *gt* (11), *hb* (12, 13), *kni* (14), and *Kr* (12–15). For the remaining genes in Table 3 we have used published consensus DNA-binding motifs as follows: *tll* (14), *ftz* (16), *cad* (17), *z* (18), and from a biochemical approach for the protein products of *Ubx* (19) and *Abd-B* (20). Table 3 summarizes the result of analyzing such consensus motifs. Contiguous oligomers of such motifs were insufficient in number to warrant plotting in Fig. 1. When ≤5 nt separates such motifs, certain of them are found to be significantly overrepresented; however, since the Markov expectation exceeds one in such cases their positions are not plotted in Fig. 1.

We also analyzed the octanucleotide, ATTTGCAT, to which the mammalian proteins OCT1 and OCT2 bind (reviewed in ref. 21). In *in vitro* cell studies, ATTTGCAT acts as a binding site for UBX and ABD-BII proteins (22). BXCALL has 78 sites for this octanucleotide, which exceeds the TMC expected number of 32.2 at a highly significant level ($<10^{-12}$) by a cumulative Poisson test.

DISCUSSION

DNA Motifs of Unknown Function. Long runs of the mononucleotide A (or its complement, T) are highly overrepresented in BXCALL (Fig. 1), as judged by the TMC expectation. The observed excessive number of runs of six or more dinucleotides of the type CA or its inverse, TG, are unlikely to be due to chance. Runs of six or more tend to occur in the introns of *Ubx*, *abd-A*, and *Abd-B* transcription units. This correlation with transcription units is consistent with the theory that such runs generate negative supercoiling during transcription (23). Runs of CT, or its inverse, AG, are also significantly overrepresented. Runs of three or more of the rotational derivative GA have been proposed to be involved in determining chromatin structure (24) and also act as enhancers of transcription when bound by the transcriptional activator known as GAGA factor (25).

Repeats of four or more CAG trinucleotides occur in the coding regions of *abd-A* and *Abd-B*. Such repeats are responsible for the long runs of glutamine in the ABD-B proteins (26) and the ABD-A proteins (27). Repeats of four or more CTG (the complement of CAG) trinucleotides occur in the noncoding region of the *Ubx* and *Abd-B* domains.

The most striking of the concatamers of hexanucleotides is that of AGATAC. Hogness *et al.* (28) reported AGATAC as a consensus sequence in the *Ubx* domain and suggested that it might serve as a binding site for "coupling proteins" that would bring distant cis-regulatory regions closer together, specifically the *bxd* and *abx/bx* regions.

Known Motifs Within the BX-C. In constructing Table 3, we have relied on consensus DNA-binding motifs that were deduced largely from *in vitro* footprinting studies. These sites are not only tentative, they are frequently degenerate and they may not be the ones used *in vivo*. Among the gap genes none except *tll* are significantly overrepresented or underrepresented in BXCALL as single motifs. Nevertheless, a function for a single Kr-binding sequence is indicated based on a mutational analysis. Specifically, two independent Hyperabdominal mutants, *Hab* and *Hab*^{2}, involve the same mutated base pair (G to A) in a Kr binding site, GGGTGAA, located at 172,672 in the *iab-2* region of the *abd-A* domain (29). These mutations are postulated to prevent binding of the Kr protein, thereby preventing the repression by that protein of *abd-A* function. The resultant overexpression of *abd-A* leads to a four-legged fly (T3 being transformed toward A1) (30).

The distribution of DNA-binding motifs for the protein products of the *Abd-B* gene is particularly interesting. TTTAT(G/T)(G/A)C, an ABD-B consensus DNA binding site, is found at four sites clustered between 43,400 and 43,821 that are located 5.8 kb 5' from the start of transcription for the mRNA that encodes ABD-BI. This number of potential binding sites and their spacing is similar to findings for autoregulatory elements of several other homeotic genes (16, 31–33).

Of special interest are clusters of the consensus binding sequence, YGAGYG (Y = T or C), of the zeste (*z*) gene product. The protein product of the wild-type *z* gene is

Table 3. Comparison of observed (O) and expected (E) numbers of DNA motifs of known function in BXCALL

Gene	Binding domain	Consensus binding site*	O	E-M[b]	E-I[c]	P[d]
Gap genes						
giant	bZip	WHWWRAYYGH	358	355.492	299.829	NS
		DCRRTYWWDW	355	354.037	299.897	NS
		$(WHWWRAYYGH)_{n=2}$	9	9.08	6.853	NS
hunchback	Zinc finger	CNYAAAAA	178	168.908	71.310	NS
		TTTTTRNG	185	161.999	68.360	0.042
		$(CNYAAAAA)_{n\geq 2}$	7	2.00	0.373	0.005
knirps	Steroid receptor	WWMTRRRHC	269	286.591	322.288	NS
		GDYYYAKWW	251	280.970	324.945	NS
		$(WWMTRRRHC)_{n=2}$	4	6.29	7.982	NS
Kruppel	Zinc finger	GGGTKAA	28	30.942	34.313	NS
		TTMACCC	41	33.094	36.142	NS
		$(GGGTKAA)_{n=2}$	1	0.11	0.095	NS
tailless	Steroid receptor	AAATTAA	171	138.840	57.709	0.005
		TTAATTT	180	143.756	57.258	0.002
		$(AAATTAA)_{n=2}$	6	1.50	0.252	0.004
Pair-rule gene						
fushi-tarazu	Homeo-domain	CCATTC	118	137.600	72.666	NS
		GAATGG	113	134.599	67.267	NS
		$(CCATTC)_{n=2}$	1	1.39	0.378	NS
Homeotic genes						
Abdominal-B	Homeo-domain	TTTATKRC	119	62.687	35.205	$<10^{-9}$
		GYMATAAA	108	63.991	34.615	$<10^{-6}$
		$(TTTATKRC)_{n=2}$	3	0.27	0.093	0.003
caudal	Homeo-domain	TTTATG	331	225.627	137.593	$<10^{-10}$
		CATAAA	343	226.846	141.718	$<10^{-12}$
		$(TTTATG)_{n\geq 2}$	19	4.20	1.494	$<10^{-6}$
Ubx	Homeo-domain	$(TAA)_4$	1	0.218	0.123	NS
		$(TTA)_4$	3	0.229	0.122	0.002
		TTAATGG	50	52.217	28.334	NS
		CCATTAA	54	53.072	29.668	NS
		$(TTAATGG)_{n=2}$	0	0.17	0.064	NS
Pairing-related gene						
zeste	—	YGAGYG	289	225.494	201.173	$<10^{-4}$
		CRCTCR	289	231.912	210.598	$<10^{-3}$
		$(YGAGYG)_{n\geq 2}$	14	4.08	3.262	$<10^{-4}$

*K = G or T; M = A or C; R = G or A; Y = C or T; N = A, G, C, or T; W = A or T; H = A, C, or T; D = A, G, or T. For each motif (m) and its complement (m'), $(m)_{n=2}$: m#m, m#m', m'#m, and m'#m', where # = spacing of ≤5 nt.
†Expectation based on the TMC method.
‡Expectation based on the independence of the single base frequencies.
§P = Probability that the observed (O) number or a larger number exceeds E-M based on the cumulative Poisson distribution. NS, not significant ($P > 0.05$).

assumed to facilitate pairing of homologous chromosome regions, whether located in cis or in trans (18). Loss of function mutants of the *z* gene suppress transvection (or pairing-dependent complementation) within the BX-C (34, 35, 43), the decapentaplegic gene (36), and the eyes-absent gene (37). Clusters of YGAGYG are not plotted in Fig. 1, since several are expected to occur by chance owing to the degeneracy of the hexanucleotide. Clusters of AGATAC have also been suggested to be involved in pairing, as already mentioned. It may be of interest that YGAGYG and AGATAC, when reduced to their purine (R)/pyrimidine (Y) hexanucleotides, YRRRYR and RRRYRY, respectively, are derivatives of one another.

The BX-C has been postulated to derive from a common ancestral sequence that tandemly duplicated and then diverged by mutation to acquire new functions (38). This postulate is consistent with the finding that the BX-C genes have their homeobox sequences highly conserved (39, 40). Whether the cis-regulatory regions will also turn out to have sequence similarities suggesting a duplication origin cannot be answered at present and will need much more extensive analysis. As a start, the distribution of significantly clustered sequences in these regions as enumerated in Fig. 1 may be viewed as an attempt to develop a "signature" for each of the cis-regulatory regions of the complex. It should be stressed that sequence analysis of the entire BX-C, as begun in this paper, will become a powerful approach to understanding regulation of the BX-C when coupled with biochemical and developmental approaches and ultimately with a comparative sequence analysis of the homologous genes in other organisms.

We thank Mary Raney, Victor Hsu, Mallory Zhang, Gretl Hornung, and John Hubenschmidt for assistance in running the computer programs. We thank Welcome Bender, Howard Lipshitz, and Joanne Topol for critical reading of the manuscript. This work was supported by research grants to E.B.L. from the March of Dimes and from the National Institutes of Health (HD06331 and HD30727).

1. Duncan, I. (1987) *Annu. Rev. Genet.* **21**, 285–319.
2. Lewis, E. B. (1992) *J. Am. Med. Assoc.* **267**, 1524–1531.
3. Martin, C. H., Mayeda, C. A., Davis, C. A., Ericsson, C. L., Knafels, J. D., Mathog, D. R., Celniker, S. E., Lewis, E. B. & Palazzolo, M. J. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 8398–8402.

4. Simon, J. (1995) *Curr. Opin. Cell Biol.* **7,** 376–385.
5. Kennison, J. A. (1993) *Trends Genet.* **9,** 75–79.
6. Weir, B. S. (1990) *Genetic Data Analysis* (Sinauer, Sunderland, MA).
7. Katz, R. W. (1981) *Technometrics* **23,** 243–249.
8. Bender, W., Akam, M., Karch, F. A., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B. & Hogness, D. S. (1983) *Science* **221,** 23–29.
9. Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M. & Lewis, E. B. (1985) *Cell* **43,** 81–96.
10. Bate, M. & Martinez-Arias, A. (1993) *The Development of Drosophila melanogaster* (Cold Spring Harbor Lab. Press, Plainview, NY).
11. Capovilla, M., Eldon, E. & Pirrotta, V. (1992) *Development (Cambridge, U.K.)* **114,** 99–112.
12. Stanojevic, D., Hoey, T. & Levine, M. (1989) *Nature (London)* **341,** 331–335.
13. Treisman, J. & Desplan, C. (1989) *Nature (London)* **341,** 335–337.
14. Hoch, M., Gerwin, N., Taubert, H. & Jackle, H. (1992) *Science* **256,** 94–97.
15. Sauer, F. & Jackle, H. (1991) *Nature (London)* **353,** 563–566.
16. Schier, A. F. & Gehring, W. J. (1992) *Nature (London)* **356,** 804–807.
17. Dearolf, C. R., Topol, J. & Parker, C. S. (1989) *Nature (London)* **341,** 340–343.
18. Benson, M. & Pirrotta, V. (1988) *EMBO J.* **7,** 3907–3915.
19. Ekker, S. C., Young, K. E., VonKessler, D. P. & Beachy, P. A. (1991) *EMBO J.* **10,** 1179–1186.
20. Ekker, S. C., Jackson, D. G., VonKessler, D. P., Sun, B. I., Young, K. E. & Beachy, P. A. (1994) *EMBO J.* **13,** 3551–3560.
21. Mitchell, P. J. & Tjian, R. (1989) *Science* **245,** 371–378.
22. Thali, M., Muller, M. M., Delorenzi, M., Matthias, P. & Bienz, M. (1988) *Nature (London)* **336,** 598–601.
23. Liu, L. F. & Wang, J. C. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7024–7027.
24. Lu, Q., Wallrath, L. L., Granok, H. & Elgin, S. C. R. (1993) *Mol. Cell. Biol.* **13,** 2802–2814.
25. Biggin, M. D. & Tjian, R. (1988) *Cell* **53,** 699–711.
26. Celniker, S. E., Keelan, D. J. & Lewis, E. B. (1989) *Genes Dev.* **3,** 1425–1437.
27. Karch, F., Bender, W. & Weiffenbach, B. (1990) *Genes Dev.* **4,** 1573–1587.
28. Hogness, D. S., Lipshitz, H. D., Beachy, P. A., Peattie, D. A., Saint, R. B., Goldschmidt-Clermont, M., Harte, P. J., Gavis, E. R. & Helfand, S. L. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50,** 181–194.
29. Shimell, M. J., Simon, J., Bender, W. & O'Connor, M. B. (1994) *Science* **264,** 968–971.
30. Lewis, E. B. (1978) *Nature (London)* **276,** 565–570.
31. Jiang, J., Hoey, T. & Levine, M. (1991) *Genes Dev.* **5,** 265–277.
32. Regulski, M., Dessain, S., McGinnis, N. & McGinnis, W. (1991) *Genes Dev.* **5,** 278–286.
33. Tremml, G. & Bienz, M. (1992) *Development (Cambridge, U.K.)* **116,** 447–456.
34. Lewis, E. B. (1954) *Am. Nat.* **88,** 225–239.
35. Lewis, E. B. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50,** 155–164.
36. Gelbart, W. M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2636–2640.
37. Leiserson, W. M., Bonini, N. M. & Benzer, S. (1994) *Genetics* **138,** 1171–1179.
38. Lewis, E. B. (1951) *Cold Spring Harbor Symp. Quant. Biol.* **16,** 159–174.
39. Scott, M. P. & Weiner, A. J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4115–4119.
40. McGinnis, W., Levine, M., Hafen, E., Kuroiwa, A. & Gehring, W. J. (1984) *Nature (London)* **308,** 428–433.
41. Genetics Computer Group (1994) *Program Manual for the Wisconsin Package, Version 8* (Genetics Computer Group, Madison, WI).
42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
43. Lindsley, D. L. & Zimm, G. G. (1992) *The Genome of Drosophila melanogaster* (Academic, San Diego).